

# Accelerated Parallel and Distributed Algorithm using Limited Internal Memory for Nonnegative Matrix Factorization

Duy Khuong Nguyen <sup>\* †</sup>      Tu-Bao Ho <sup>\* ‡</sup>

July 1, 2015

## Abstract

Nonnegative matrix factorization (NMF) is a powerful technique for dimension reduction, extracting latent factors and learning part-based representation. For large datasets, NMF performance depends on some major issues: fast algorithms, fully parallel distributed feasibility and limited internal memory. This research aims to design a fast fully parallel and distributed algorithm using limited internal memory to reach high NMF performance for large datasets. In particular, we propose a flexible accelerated algorithm for NMF with all its  $L_1$   $L_2$  regularized variants based on full decomposition, which is a combination of an anti-lopsided algorithm and a fast block coordinate descent algorithm. The proposed algorithm takes advantages of both these algorithms to achieve a linear convergence rate of  $\mathcal{O}(1 - \frac{1}{\|Q\|_2})^k$  in optimizing each factor matrix when fixing the other factor one in the sub-space of passive variables, where  $r$  is the number of latent components; where  $\sqrt{r} \leq \|Q\|_2 \leq r$ . In addition, the algorithm can exploit the data sparseness to run on large datasets with limited internal memory of machines. Furthermore, our experimental results are highly competitive with 7 state-of-the-art methods about three significant aspects of convergence, optimality and average of the iteration number. Therefore, the proposed algorithm is superior to fast block coordinate descent methods and accelerated methods.

**Keywords:** Non-negative matrix factorization, Anti-lopsided algorithm, Coordinate descent algorithm, and Parallel and distributed algorithm.

---

<sup>\*</sup>Japan Advanced Institute of Science and Technology, Japan

<sup>†</sup>University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam

<sup>‡</sup>John von Neumann Institute, Vietnam National University, Ho Chi Minh City, Vietnam

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Background and Related Works</b>	<b>4</b>
2.1	Background . . . . .	4
2.2	Related works . . . . .	5
<b>3</b>	<b>Proposed Algorithm</b>	<b>7</b>
3.1	Iterative multiplicative update accelerated algorithm . . . . .	7
3.2	Full decomposition for NMF . . . . .	8
3.3	Parallel and distributed algorithm using limited internal memory . . . . .	8
3.4	Fast algorithm for nonnegative quadratic programming . . . . .	10
3.5	Extensions for $L_1$ $L_2$ regularized NMF . . . . .	14
<b>4</b>	<b>Theoretical Analysis</b>	<b>15</b>
4.1	Convergence . . . . .	15
4.2	Complexity . . . . .	16
<b>5</b>	<b>Experimental evaluation</b>	<b>17</b>
5.1	Convergence . . . . .	18
5.2	Optimality . . . . .	19
5.3	Average of iteration number . . . . .	20
5.4	Running on large datasets . . . . .	20
5.5	Regularized NMF extensions . . . . .	21
<b>6</b>	<b>Conclusion</b>	<b>21</b>

# 1 Introduction

Nonnegative matrix factorization (NMF) is a powerful technique widely used in applications of data mining, signal processing, computer vision, bioinformatics, etc. [Zha11b]. Fundamentally, NMF has two main purposes. First, it reduces dimension of data making learning algorithms faster and more effective as they often work less effectively due to the curse of dimensionality [HV05]. Second, NMF helps extracting latent components and learning part-based representation, which are the significant distinction from other dimension reduction methods such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), Vector Quantization (VQ), etc. This feature originates from transforming data into lower dimension of latent components and non-negativity constraints [DS04, Gil14, LS+99].

In the last decade of fast development, there were remarkable milestones. The two first milestones in early days of the NMF historical development were its mathematical formulations as positive matrix factorization with Byzantine algorithms [PT94] and as parts-based representation with a simple effective algorithm [LS+99]. The last decade has witnessed the rapid NMF development [Zha11b, WZ13]. Various works on NMF can be viewed in three major perspectives: variants of NMF, algorithms and applications. In particular, variants of NMF are based on either divergence functions [SL01, Zha11a], or constraints [Hoy04, PMCK+06], or regularizations [Cho08, LAW+07]. Most NMF algorithms were developed along two main directions: geometric greedy algorithms [TKWB11] and iterative multiplicative update algorithms. Although geometric greedy algorithms are usually fast, they are hard to trade off complexity, optimality, loss information and sparseness.

More recently, it is well recognized that the most challenging problems in iterative multiplicative update algorithms for NMF are fast learning, limited internal memory, parallel distributed computation, among others. In particular, fast learning is essential in learning NMF models from large datasets, and it is indeed difficult to carry out them when the number of variables is very large. In addition, the limited internal memory is one of the most challenging requirements for big data [GWLT13], because data has been exploring rapidly while the internal memory of nodes is always limited. Finally, parallel and distributed computation makes NMF applications feasible for big data [LYF+10].

To deal with these challenges, this work develops an accelerated algorithm for NMF and its  $L_1$   $L_2$  regularized variants having several major advantages that are summarized in Table 1. In this paper, we contribute five folders as follows:

- *NMF and its variants*: We fully decompose NMF and its  $L_1$   $L_2$  regularized variants into non-negative quadratic programming problems. This decomposition makes the proposed algorithm flexible to adapt all  $L_1$   $L_2$  regularized NMF in an unified framework that can trade-off the quality of information loss, sparsity and smoothness.

- *Algorithm*: We employ a combinational algorithm of an anti-lopsided algorithm and a fast block coordinate descent algorithm for non-negative quadratic programming. The algorithm reduces variable scaling problems to achieve linear convergence rate of  $(1 - \frac{1}{\|Q\|_2})^k$  in optimizing each factor matrix in the sub-space of passive variables, which is advanced to fast coordinate methods and accelerated methods in terms of efficiency as well as convergence rate. In addition, the size of optimization problem is reduced into  $r$  ( $r \ll m, n$ ), which is the

Table 1: Comparison Summary of NMF solvers

Criteria	Inexact		Exact				Accelerated			
	MUR	PrG	Qn	Nt	AcS	BIP	FCD	AcH	Ne	Alo
Guaranteed Convergence	✗	✗	✗	✗	✗	✗	✗	✗	✓ $\frac{1}{k^2}$	✓ $(1 - \frac{1}{\ Q\ _2})^k$
Exploit Data Sparseness	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓
Limited Internal Memory	$\mathcal{O}(mn + r(r + n + m))$									$\mathcal{O}(r(r + n))$
Fully Parallel & Distributed	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Optimization Problem Size	$r(m, n)$		$r$		$r(n, m)$		$(m, n)r(m, n)$		$r$	

✓ means considered, and ✗ means not considered

$\sqrt{r} \leq \|Q\|_2 \leq r$ ,  $n \times m$  is the data matrix size,  $r$  is the number of latent components

$(m, n) = \max(m, n)$ , and  $r(m, n) = r \cdot \max(m, n)$

Abbreviations: **MUR**: Multiplicative Update Rule [LS<sup>+</sup>99]; **PrG**: Projected Gradient methods [Lin07b]; **Nt**: Newton-type methods [KSD07]; **Qn**: Projected Quasi-Newton [ZC06]; **AcS**: Fast Active-set-like method [KP08a]; **BIP**: Block Principal Pivoting method [KP08b]; **FCD**: Fast Coordinate Descent methods with variable selection [HD11]; **AcH**: Accelerated Hierarchical Alternating Least Squares [GG12]; **Ne**: Nesterov’s optimal gradient method [GTLY12]; **Alo**: The proposed method.

smallest among the state-of-the-art methods. Hence, the algorithm has the low complexity and converges very fast to the optimal solution, and it is highly potential to be applied in alternating least squares methods for factorization models.

- *Parallel and Distribution*: The proposed algorithms are fully parallel and distributed on limited internal memory systems, which is crucial for big data when computing nodes having limited internal memory that cannot hold the whole dataset.

- *Implementation*: The proposed algorithms are convenient to implement for hybrid multi-core distributed systems because this algorithm works on each individual instance and each latent feature.

- *Comparison*: This is the first time that state-of-the-art algorithms in different research directions for NMF are compared together.

The rest of paper is organized as follows: Section 2 discusses the background and related works of NMF; Section 3 mentions our proposed algorithm; Section 4 gives a complexity analysis of our proposed algorithms; Section 5 experimentally compares our proposed algorithm with state-of-the-art algorithms for NMF among remarkable approaches; our conclusion is stated in Section 6.

## 2 Background and Related Works

### 2.1 Background

Mathematically, NMF in Frobenius norm is defined as follows:

**Definition 1 [NMF]**: Given a dataset consisting of  $m$  vectors in a  $n$ -dimension space  $V = [V_1, V_2, \dots, V_m] \in R_+^{n \times m}$ , where each vector presents a data instance. NMF seeks to

decompose  $V$  into a product of two nonnegative factorizing matrices  $G$  and  $F$ , where  $G = [G_1, \dots, G_r] \in R_+^{n \times r}$  and  $F = [F_1, \dots, F_m] \in R_+^{r \times m}$  are the latent component matrix and the coefficient matrix respectively,  $V \approx GF$ , in which the quality of approximation can be guaranteed by the objective function in Frobenius norm:  $D(V||GF) = \|V - GF\|_2^2$ .

Although NMF is a non-convex problem, optimizing each factor matrix when fixing the other one is a convex problem. In other words,  $F$  can be traced when  $G$  is fixed, and vice versa. Furthermore,  $F$  and  $G$  have different roles although they are symmetric in the objective function.  $G$  are latent components to represent data instances  $V$  by coefficients  $F$ . Hence, NMF can be considered as a latent factor model of latent components  $G$ , and learning this model is equivalent to find out latent components  $G$ . Therefore, in this paper, we propose an accelerated parallel and distributed algorithm to learn NMF models  $G$  for large datasets.

## 2.2 Related works

NMF algorithms can be divided into two groups: the greedy algorithms and the iterative multiplicative update algorithms. The greedy algorithms [TKWB11] are often based on geometric interpretability, and they can be extremely fast to deal with large datasets. However, it is hard to trade off complexity, optimality, loss information and sparseness. The iterative multiplicative update algorithms such as “two-block coordinate descent” often consist of two steps, each of them fixes one of two matrices to replace the other matrix for obtaining the convergence of the objective function. There are numerous studies on these algorithms, see Table 1, because NMF is nonconvex, though two steps corresponding to two non-negative least square (NNLS) sub-problems are convex [GTLY12, KHP14]. In addition, various constraints and optimization strategies have been used to trade off the convexity, information loss, complexity, sparsity, and numerical instability.

Based on the optimization updating strategy, these iterative multiplicative update algorithms can be further divided into three sub-groups:

- *Inexact Block Coordinate Descent*: The algorithms’ common characteristic is their usage of gradient methods to seek an approximate solution for NNLS problems, which is neither optimal nor fulfilling of fast approximations and accelerated conditions. Lee *et al.* [LS<sup>+</sup>99] proposed the (basic) NMF problem and simple multiplicative updating rule (MUR) algorithm using first-order gradient method to learn the part-based representation. Seung *et al.* [SL01] concerned rescaling gradient factors with carefully selected learning rate to achieve a faster convergence rate. Subsequently, Lin [Lin07a] modified MUR, which is theoretically proved getting a stationary point (a local minimum optimization). However, that algorithm cannot improve the convergence rate. Berry *et al.* [BBL<sup>+</sup>07] projected nonnegative least square (PNLS) solutions into nonnegative quadratic space by setting negative entries in the matrices to zero. Although this algorithm does not guarantee the convergence, it is widely applied in real applications. In addition, Bonettini *et al.* [Bon11] used line search based on Amijo rule to obtain better solutions for matrices. Theoretically, this method can achieve optimal solutions for factor matrices as exact block coordinate descent group, but it very slowly tends to stationary points because the line search is time-consuming.

- *Exact Block Coordinate Descent*: In contrast to the first sub-group, the common characteristic in this group is obtaining optimal solutions for two NNLS problems in each iteration. Zdunek *et al.* [ZC06] employed second-order quasi-Newton method with inverse of Hessian matrix to estimate the step size, aiming to a faster convergence than projected methods. However, this algorithm may be slow and non-stable because of the line search. Subsequently, Kim *et al.* [KSD07] used rank-one to Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm to approximate the inverse of Hessian matrix. Furthermore, Chih-Jen Lin [Lin07b] proposed several algorithms based on projected gradient methods and exact line search. Theoretically, this method can obtain more accurate solutions, however it is time-consuming because of exact line search and the number of iterations increased by the large number of variables. Moreover, Kim *et al.* [KP08a, KP08b] proposed two active-set methods based on Karush-Kuhn-Tucker (KKT) conditions, in which the variables are divided into two sets: a free set and an active set. Only the free set contains variables that can optimize the objective functions. Removing the number of redundant variables makes their algorithms improve the convergence rate significantly. However, the method still has heavy computation for large-scale problems.

- *Accelerated Block Coordinate Descent*: The accelerated methods use fast solution approximations satisfying accelerated conditions to reduce the complexity and to keep fast convergence. The accelerated conditions are different constraints in different methods to guarantee convergence to the optimal solution in comparison with the initial value. These accelerated methods are developed due to the limitation of inexact methods having slow convergence, and exact methods having high complexity in each iteration. Particularly, for inexact methods, they have slow convergence because of the high complexity of solution approximations in each iteration or a large number of iterations that leads to the highly expensive computation between two sequential iterations. Furthermore, the exact methods have high complexity in each iteration, however obtaining optimal solutions in every iteration is controversial because it can lead to zig-zag problems when optimizing a non-convex function of two independent sets of variables.

Firstly, Hsieh *et al.* [HD11] proposed a fast coordinate descent method with the best variable selection to reduce the objective function. The algorithm iteratively selects variables to update the approximate solution until the accelerated stopping condition  $\max_{ij} D_{ij}^G < \epsilon p_{\text{init}}$  satisfied, where  $D_{ij}^G$  is the reduction of the objective function based on the variable  $G_{ij}$ , and  $p_{\text{init}}$  is the maximum initial reduction over the matrix  $G$ . Although the greedy update method does not have guaranteed convergence, it has the fast convergence speed in many reports.

Subsequently, Gillis and Glineur [GG12] proposed a number of accelerated algorithms using fast approximation by fixing all variables but excepting a single column of factor matrices. This framework improved significantly the effectiveness of multiplicative updates [LS01], hierarchical alternating least squares (HALS) algorithms [CZA07] and projected gradients [Lin07b]. These algorithms achieve the accelerated condition in each iteration such as that  $\|G^{(k,l+1)} - G^{(k,l)}\|_2^2 \leq \epsilon \|G^{(k,1)} - G^{(k,0)}\|_2^2$  is the stopping condition when optimizing the objective function on  $G$  if fixing  $F$ . Although these greedy algorithms does not have guaranteed convergence, their results are highly competitive with the inexact and exact methods.

More recently, Guan *et al.* [GTLY12] employed Nesterov’s optimal methods to optimize NNLS with fast convergence rate  $\mathcal{O}(1/k^2)$  to achieve the accelerated convergence condition  $\|\frac{\partial f}{\partial G_{(k,l+1)}}\|_2^2 \leq \epsilon \|\frac{\partial f}{\partial G_{(k,0)}}\|_2^2$ . Although Guan *et al.*’s method [GTLY12] has a fast convergence rate  $\mathcal{O}(1/k^2)$ , it has several drawbacks such as working on the whole factor matrices, and less flexibility for regularized NMF variants. Furthermore, this approach does not consider the issues of parallel and distribution, and they require numerous iterations to satisfy the accelerated condition because the step size is limited by  $\frac{1}{L}$ , where  $L$  is Lipschitz constant.

To deal with the above issues of accelerated methods, in next section, we propose an accelerated parallel and distributed algorithm for NMF and its regularized  $L_1$   $L_2$  variants with linear convergence in optimizing each factor matrix when fixing the other factor one.

### 3 Proposed Algorithm

To read easily, this section hierarchically presents our proposed algorithm. First, an iterative multiplicative update accelerated algorithm is introduced. Then, a transformational technique fully decomposes the objective functions of NMF into basic computation units as nonnegative quadratic programming (NQP) problems. After that, a modified version of the algorithm is proposed to deal with the issues of parallel and distributed systems. Subsequently, a combinational method of an anti-lopsided algorithm and a fast coordinate descent algorithm is developed to effectively solve NQP problems. Finally, extensions for  $L_1 L_2$  regularized NMF is discussed.

#### 3.1 Iterative multiplicative update accelerated algorithm

For solving NMF, we employ an iterative multiplicative update accelerated algorithm, like expectation-maximization (EM) algorithm, presented in Algorithm 1. This algorithm consists of two main steps: one for finding  $F^+$  ( $F^+$  is updated  $F$  in the iteration) when fixing  $G$  and the other for finding  $G^+$  when fixing  $F$ . In the first step called *E-step*, we find  $F^+$ , each column of which  $F_i^+$  is the new representation of a data instance  $V_i$  in the new space of latent components  $G$ . Meanwhile, the other one, called *M-step*, learns new latent components.

---

#### Algorithm 1: Iterative Multiplicative Update Accelerated Algorithm

---

**Input:** Data matrix  $V = \{V_i\}_{i=1}^m \in R_+^{n \times m}$  and  $r$ .  
**Output:** Latent components  $G = \{G_k\}_{k=1}^r$ .

- 1 **begin**
- 2     Randomize  $r$  nonnegative latent components  $\in R_+^{n \times r}$  ;
- 3     **repeat**
- 4         **E-step:** Fixing  $G$  to find  $F^+$  such that the accelerated condition is satisfied;
- 5         **M-step:** Fixing  $F$  to find  $G^+$  such that the accelerated condition is satisfied;
- 6     **until** *Convergence condition is satisfied*;

---



### 3.2 Full decomposition for NMF

This section discusses decomposing the objective function of NMF into non-negative quadratic programming (NQP) problems, which aims to fully parallelize and distribute the NMF computation. Particularly, in Algorithm 1, the E-step is to find new coordinates of data instances in the space of latent components  $G$  by minimizing  $J(V||GF) = \|V - GF\|_2^2 = \sum_{i=1}^m \|V_i - GF_i\|_2^2$ . Hence, minimizing  $J(V||GF)$  is equivalent to independently minimizing  $\|V_i - GF_i\|_2^2$  for each instance  $i$  since  $G$  is fixed. Similarly, the M-step is also equivalent to independently minimizing  $\|V_j^T - F^T G_j^T\|_2^2$  for each feature  $j$ , where  $F$  is fixed. Hence, the basic computation units are nonnegative least-squares (NNLS) problems [LH74].

For large datasets  $n, m \gg r$ , we equivalently turn these problems into nonnegative quadratic programmings (NQP):

$$\begin{aligned} & \underset{x}{\text{minimize}} && \frac{1}{2} \|Ax - b\|_2^2 \\ & \text{subject to} && x \succeq 0 \in R^r \\ & \text{where} && A \in R_+^{nr}, b \in R_+^n \end{aligned} \tag{1}$$

equivalent to

$$\begin{aligned} & \underset{x}{\text{minimize}} && f(x) = \frac{1}{2} x^T H x + h^T x \\ & \text{subject to} && x \succeq 0. \\ & \text{where} && H = A^T A, h = -A^T b \end{aligned} \tag{2}$$

Hence, finding new coefficients  $F^+$  and new latent components  $G^+$  can be fully paralleled and distributed into basic computation units as solving NQP problems.

### 3.3 Parallel and distributed algorithm using limited internal memory

In this section, we design a parallel and distributed algorithm using limited internal memory for learning NMF model  $G$ , see Fig. 1, which is a modified version of Algorithm 1.

For large datasets, the computation can be untimely performed in a single process, so parallel and distributed algorithm environments are employed to speed up the computation. For parallel and distributed systems, we often face two major issues: dependency of computation units and limited internal memory computing nodes. In particular, computation units must be independently conducted as much as possible, since any dependency of computing elements will increase the complexity of implementation and the delay of data transfer over the network that reduces the performance of system. Furthermore, for these parallel distributed systems, computation units are executed on computing nodes within a limited internal memory. In addition, accessing external memory will increase the complexity and reduce the performance.

For our proposed approach, the computation can be fully paralleled and distributed, and use limited internal memory in computing nodes because the objective function is properly



---

**Algorithm 2:** Parallel and Distributed Algorithm

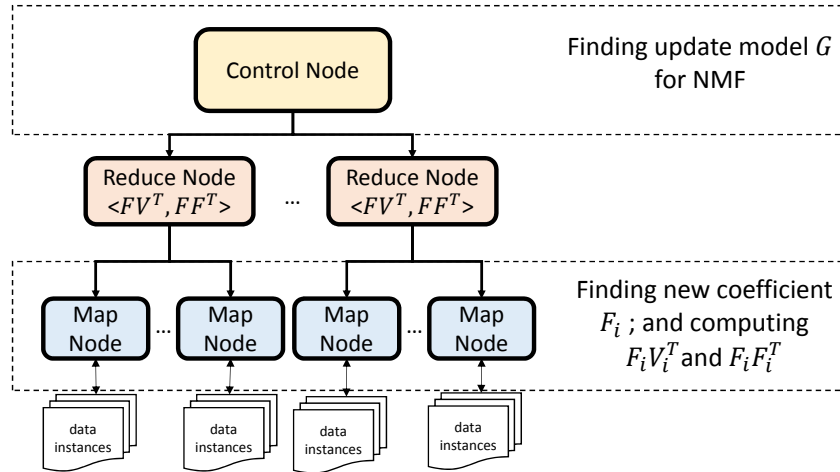
---

**Input:** Data matrix  $V = \{V_m\}_{m=1}^m \in R_+^{nm}$  and  $r$ .

**Output:** Latent components  $G = \{g_k\}_{k=1}^r$ .

```
1 begin
2   Randomize  $r$  nonnegative latent components  $G \in R_+^{nr}$ ;
3   repeat
4      $Y = 0 \in R^{nr}$  /*  $Y = FV^T$  */;
5      $H = 0 \in R^{rr}$  /*  $H = FF^T$  */;
6      $Q = GG^T$ ;
7      $maxStop = 0$ ;
8     /*Parallel and distributed*/
9     for  $i = 1$  to  $m$  do
10      /*call Algorithm 3*/;
11       $F_i = \underset{x \in R^r \succeq 0}{\text{Minimizing}} (x^T Q x / 2 - V_m^T G^T x)$ ;
12       $Y = Y + F_i V_i^T$ ;
13       $H = H + F_i F_i^T$ ;
14    /*Parallel and distributed*/
15    for  $j = 1$  to  $n$  do
16      /*call Algorithm 3*/;
17       $G_j = \underset{x \in R^r \succeq 0}{\text{Minimizing}} (x^T H x / 2 - Y_n^T x)$ ;
18  until Convergence condition is satisfied;
```

---



**Figure 1:** Distributed System Diagram for NMF

decomposed to NQP problems. Particularly, Algorithm 2 presents a modified version of iterative multiplicative update algorithms, in which computation units are fully paralleled and distributed. In addition,  $Q = GG^T$  is precomputed to reduce the complexity, and finding new coefficients  $F_i$  can be independently computed and distributed. Remarkably, the most heavy computation of  $Y = FV^T$  and  $H = FF^T$  is divided into computing  $F_iV_i^T$  and  $F_iF_i^T$  to be parallel and distributed.

Particularly, the distributed system using MapReduce is described in Fig. 1. In this computing model, data instances and the instance projection are parallel and distributed over the Map nodes. The Reduce nodes sum up the results  $F_iF_i^T$  and  $F_iV_i^T$  of the Map nodes. Subsequently, in the M-step, the results  $FF^T$  and  $FV^T$  are employed to compute latent components  $G$ . This M-step computation can be conducted by a single machine or a distributed system, which depends on the dimension of problem because the time to distribute this computation over the network is usually considerable.

In comparison with the previous algorithms, this computing model is much more effective than the previous models [GNHS11, LYF<sup>+</sup>10, SLR10] by the following reasons:

- The necessary memory used in computing nodes is  $\mathcal{O}(\text{size}(G, Y, H)) = \mathcal{O}(r(r + n))$ . The necessary memory used in the controlled node is  $\mathcal{O}(\text{size}(G, Y, H, Q)) = \mathcal{O}(r(r + n))$ . In practice, approximate solutions of NQP problems should be cached in hard disks in order to increase accuracy and reduce the number of iterations.
- At each distributed iteration, the computation is fully decomposed into basic computations units, which enhances the convergence speed to the optimal solution because the size of optimization is significantly reduced. Furthermore, the expensive computation  $FV^T$  and  $FF^T$  is fully parallelized and distributed over the computing nodes.
- The computational model is conveniently implemented because computing NMF model is divided into basic computation units as NQP problems that are independently solved, and the optimization is carried out on vectors instead of matrices.

In the next section, we propose a novel algorithm, Algorithm 3, to solve approximately NQP problems, which is robust and effective because it only uses the first derivative and does not consider the ill-condition of matrix inverse.

### 3.4 Fast algorithm for nonnegative quadratic programming

In this section, we briefly review the literature before proposing the novel algorithm to solve NQP Problem 2 for real large-scale NMF applications.

Regarding algorithms for NNLS and its equivalent problem NQP, numerous algorithms are proposed to deal with high dimension [CP09]. Generally, methods for solving NNLS can be divided into two groups: active-set and iterative methods [CP09]. Active-set methods are traditional to solve accurately [BDJ97, LH74]. However, they require heavy computation in repeatedly computing  $(A^T A)^{-1}$  with different set of passive variables. Hence, iterative methods that can handle multiple active constraints in each iteration have more potential for fast NMF algorithms [CP09, KSD06, KDD13]. Hence, iterative methods can deal with more large-scale problems. Among the fast iterative methods, the coordinate descent method [FHN05] has fast approximation, but has the zip-zag problem when the solution

requires high accuracy. In addition, accelerated methods [Nes83] has a fast convergence  $\mathcal{O}(1/k^2)$  [GTLY12], which only require the first order derivative. However, one major disadvantage of the methods is that they require a big number of iterations because their step size is limited by  $\frac{1}{M}$  that can be very small for large-scale problems; where  $M$  is Lipschitz constant. More recently, the anti-lopsided algorithm [NH15] re-scale variables to obtain a linear convergence in the sub-space of passive variables. Unfortunately, the passive variables are unknown in advance, so several iterations are required to determine them. In addition, the complexity of each iteration is considerable about  $\mathcal{O}(r^2)$ .

Therefore, we propose a combinational algorithm of the anti-lopsided algorithm [NH15] and the greedy coordinate block descent algorithm [HD11] to reduce the number of iterations as well as complexity. Particularly, the proposed algorithm, Algorithm 3, contains two main steps: The first step, from Line 4 to Line 7, rescales variables to avoid rescaling problems of the first order methods by replacing  $y = x \cdot \sqrt{\text{diag}(H)}$ , we have:

$$f(x) = \frac{1}{2}x^T H x + h^T x = \frac{1}{2}y^T Q y + q^T y \quad (3)$$

where  $Q = \frac{H}{\sqrt{\text{diag}(H)\text{diag}(H)^T}}$  and  $q = \frac{h}{\sqrt{\text{diag}(H)}}$  such that  $\frac{\partial^2 f}{\partial^2 y_i} = Q_{ii} = \frac{H_{ii}}{\sqrt{H_{ii}H_{ii}}} = 1$  for  $\forall i$ . By the way, the rate of change of a quantity through variables equals to a constant and the exact line search can converge at an exponential rate of  $(1 - \frac{1}{\|Q\|_2})^r$  [NH15] in the sub-space of passive variables. The passive variables are variables belongs the set  $P = \{x_i | x_i > 0 \text{ or } \nabla f_i < 0\}$  that changes through iterations.

The second step contains a loop of iterations, from Line 9 to Line 21, each of which is clearly divided into two parts: one part from Line 11 to Line 14 inherited from the anti-lopsided algorithm [NH15] and the other from Line 16 to Line 20 based on the fast coordinate descent algorithm [HD11]. The anti-lopsided algorithm guarantees the linear convergence  $(1 - \frac{1}{\|Q\|_2^2})^k$  ( $\|Q\|_2^2 \leq r$ ) in the sub-space of passive variables to avoid the zip-zag problem of the fast coordinate descent algorithm, while the coordinate block descent algorithm speeds up the convergence to the final optimal set of passive variables. In addition, the complexity of each part is still kept in  $\mathcal{O}(r^2)$ . As a result, the proposed algorithm will utilize advantages of both algorithms to attain a fast convergence, while retaining the same low complexity  $\mathcal{O}(r^2)$  of each iteration.

To comprehend the proposed algorithm's effectiveness, we consider optimizing Function 4:

$$f(x) = \frac{1}{2}x^T \begin{bmatrix} 1 & 0.1 \\ 0.1 & 10 \end{bmatrix} x + [-80 \ -100]x \quad (4)$$

The exact search gradient algorithm, from Line 11 to Line 14, starting with  $x_0 = [200 \ 20]^T$  performs 59 iterations to reach the optimal solution, see Fig. 2. However, the proposed algorithm only needs 1 iterations to reach the optimal solution, see Fig. 3 because we optimize Function 5 instead of Function 4; where Function 5 is equivalently obtained by applying the steps from Line 11 to Line 14. The exact search gradient algorithm becomes much faster because the shape of Function 5 become more sphere, and its derivative is more effective to

optimize the objective function.

$$f(y) = \frac{1}{2}y^T \begin{bmatrix} 1 & \frac{0.1}{\sqrt{10}} \\ \frac{0.1}{\sqrt{10}} & 1 \end{bmatrix} y + \begin{bmatrix} -80 & -100 \\ \sqrt{10} & \sqrt{10} \end{bmatrix} y \quad (5)$$

---

**Algorithm 3:** Fast Combinational Algorithm for NQP

---

**Input:**  $H \in R^{r \times r}$  and  $h \in R^r$  and  $x_0$

**Output:**  $x$  minimizing  $\frac{1}{2}x^T Hx + h^T x$   
subject to:  $x \succeq 0$

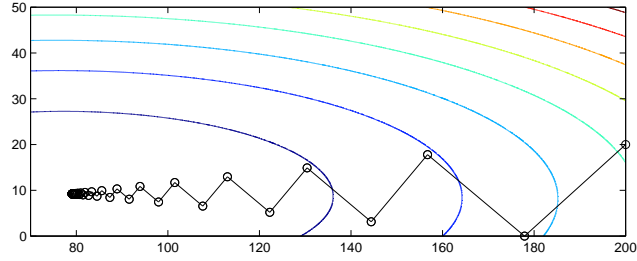
```

1 begin
2   /*Having a variable  $maxStop = 0$  for each thread of computation */;
3   /*Re-scaling variables*/;
4    $Q = \frac{H}{\sqrt{diag(H)diag(H)^T}}$ ;
5    $q = \frac{h}{\sqrt{diag(H)}}$ ;
6   /*Solving NQP: minimizing  $f(x) = \frac{1}{2}x^T Qx + q^T x^*$ */;
7    $x = x_0 \cdot \sqrt{diag(H)}$ ;
8    $\nabla f = Qx + q$ ;
9   repeat
10    /*Exact Line Search*/;
11     $\nabla \tilde{f} = \nabla f[x > 0 \text{ or } \nabla f < 0]$ ;
12     $\alpha = \operatorname{argmin} \alpha f(x_k - \alpha \nabla \tilde{f}) = \frac{\|\nabla \tilde{f}\|_2^2}{\nabla \tilde{f}^T Q \nabla \tilde{f}}$ ;
13     $x_k = [x_{k-1} - \alpha \nabla \tilde{f}]_+$ ;
14     $\nabla f_k = \nabla f_k + Q(x_k - x_{k-1})$ ;
15    /*Block Coordinate Descent*/;
16    for  $t=1$  to  $n$  do
17       $\Delta x_i = \max(0, [x_k]_i - \frac{f_i}{Q_{ii}}) - [x_k]_i \forall i$ ;
18       $p = \operatorname{argmax} i |f(x_k) - f(x_k + \Delta x_i)|$ ;
19       $\nabla f_k = \nabla f_k + Q_p \Delta x_p$ ;
20       $[x_k]_p = [x_k]_p + \Delta x_p$ ;
21    until  $(\|\tilde{f}_k\|_2^2 \leq \epsilon \|\tilde{f}_0\|_2^2) \text{ or } (\|\tilde{f}_k\|_2^2 \leq maxStop)$ ;
22     $maxStop = \max(maxStop, \|\tilde{f}_k\|_2^2)$ ;
23    return  $\frac{x_k}{\sqrt{diag(H)}}$ 

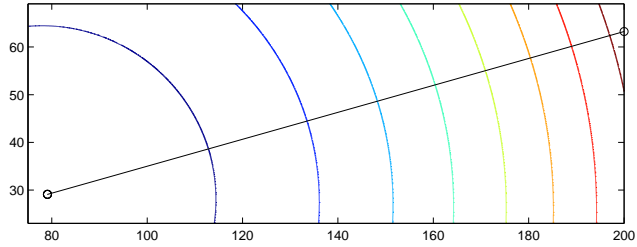
```

---

Moreover, Algorithm 3 only attains approximate solutions because achieving the optimal solution is controversial for the reasons that its computation is expensive and it can leads to the zig-zag problem in optimizing a non-convex function. In addition, it is necessary to control and balance the quality of the convergence to the optimal solution. Hence, we employ an accelerated condition  $(\|\tilde{f}_k\|_2^2 \leq \epsilon \|\tilde{f}_0\|_2^2)$  to regulate the quality of the convergence to the



**Figure 2:** 59 optimizing steps in iterative exact line search method using the first order derivative for the function 4 starting at  $x_0 = [200 \ 20]^T$



**Figure 3:** 1 optimizing steps in iterative exact line search method using the first order derivative for the function 5 starting at  $y_0 = x_0 \sqrt{\text{diag}(H)}$

optimal solutions of the NQP problems in comparison with initial values and a fast-break condition ( $\|\tilde{f}_k\|_2^2 \leq \text{maxStop}$ ) to balance the quality of the convergence among variables in each thread of the computation. As a result, the objective function converges faster through iterations; and the complexity and average of the iteration number are reduced significantly.

### 3.5 Extensions for $L_1$ $L_2$ regularized NMF

In this section, we consider solutions for  $L_1$   $L_2$  regularized NMF variants to control the quality of NMF.  $L_1$  regularized NMF [Hoy02] aims to achieve sparse solutions in optimization problems. Usually, only the coefficient matrix  $F$  is penalized to control its sparsity. Meanwhile, concerning  $L_2$  regularized NMF, the penalty terms of  $F$  and  $G$  are added to control smoothness of solutions in NMF [PPP06]. Fortunately, the objective functions of  $L_1$   $L_2$  regularized NMF can be turned into NQP problems, of which solutions are completely similar to the general NMF. Particularly, in the most general variant, the objective function  $J(X||GF)$  is formulated by:

$$\|X - GF\|_2^2 + \mu_1\|F\|_1 + \beta_1\|G\|_1 + \mu_2\|F\|_2^2 + \beta_2\|G\|_2^2 \quad (6)$$

where  $\|\cdot\|_1$  is the  $L_1$ -norm,  $\|\cdot\|_2$  is the  $L_2$ -norm, and  $\mu_1, \mu_2, \beta_1, \beta_2$  are regularized parameters that tradeoff the sparsity and the smoothness of the information loss. Obviously, both the E-step and the M-step need to solve the same NNLS problems when one of the two matrices is fixed. For example, in a E-step, we can minimize the objective function by independently solving NQP problems when fixing  $G$ :

$$\begin{aligned} J(X||GF) &= \frac{1}{2}\|X - GF\|_2^2 + \mu_1\|F\|_1 + \mu_2\|F\|_2^2 + C \\ &= \sum_{m=1}^M \left( \frac{1}{2}\|X_m - GF_m\|_2^2 + \mu_1(1^K)^T F_m + \mu_2 F_m^T I F_m \right) + C \\ &= \sum_{m=1}^M \left( \frac{1}{2} F_m^T Q F_m + q^T F_m \right) + C \end{aligned} \quad (7)$$

where  $Q = G^T G + 2\mu_1 I$ ,  $q^T = -X_m^T G^T + \mu_2 1^K$  and  $C$  is a constant.

This transformation from minimizing the objective functions into solving NQP problems independently is comprehensive to understand and simplify the variants of NMF problems as much as possible. As a result, we can conveniently implement NMF and its  $L_1$   $L_2$  regularized variants in parallel distributed systems as in sub-section 3.3.

In comparison with the previous algorithms that optimizing the objective function works on the whole of matrices, this approach decomposing the objective function is easier to parallelize and distribute the computation. Additionally, it is faster to reach the solutions because it only performs on a smaller set of variables.

## 4 Theoretical Analysis

In this section, we investigate the convergence of Algorithm 3 and the complexity of Algorithm 2 using Algorithm 3

### 4.1 Convergence

In this section, we only consider the convergence rate of Algorithm 3 by the general NMF for the two following reasons. Firstly,  $L_1$  regularized coefficients do not affect on the complexity. Secondly,  $L_2$  regularized coefficients are often small, and they change Lipschitz constants  $m$  and  $M$  by adding a small positive value, where  $m$  and  $M$  are positive Lipschitz constants of strongly convex function  $f(x)$  satisfying  $mI \preceq \frac{\partial^2 f}{\partial^2 x} \preceq MI$  and  $I$  is the identity matrix. Hence,  $L_2$  regularized coefficients slightly change the convergence rate because it depends on  $\frac{m}{M}$ .

Based on [NH15], consider the complexity of Algorithm 3, we have:

**Theorem 4.1.** Algorithm 3 linearly converges at the rate of  $\mathcal{O}(1 - \frac{1}{\|Q\|_2})^k$  in the sub-space of passive variables, where  $\sqrt{r} \leq \|Q\|_2 \leq r$ ,  $r$  is the dimension of solutions or the number of latent factors, and  $k$  is the number of iterations.

*Proof.* From [NH15], we have:

**Remark 1:** After  $(k + 1)$  iterations,  $f(x^{k+1}) - f^* \leq (1 - \frac{m}{M})^k (f(x^0) - f^*)$ , where  $mI \preceq \nabla^2 f \preceq MI$ ,  $f^*$  is the minimum value of  $f(x)$ , and  $f(x)$  is a strongly convex function of the passive variables.

We have  $\nabla^2 f = Q$ , and

$$x^T I x \leq \sum_{i=1}^r \sum_{j=1}^r Q_{ij} x_i x_j = x^T Q x \text{ since } x \geq 0, Q \geq 0, \text{ and } Q_{ii} = 1. \Rightarrow I \preceq Q.$$

Moreover, based on Cauchy-Schwarz inequality, we have:

$$\begin{aligned} \left( \sum_{i=1}^r \sum_{j=1}^r Q_{ij} x_i x_j \right)^2 &\leq \left( \sum_{i=1}^r \sum_{j=1}^r Q_{ij}^2 \right) \left( \sum_{i=1}^r \sum_{j=1}^r (x_i x_j)^2 \right) \\ &\Rightarrow \sum_{i=1}^r \sum_{j=1}^r Q_{ij} x_i x_j \leq \sqrt{\|Q\|_2^2 \left( \sum_{i=1}^r x_i^2 \right)^2} \\ &\Leftrightarrow x^T Q x \leq \|Q\|_2 x^T I x \quad (\forall x) \quad \Leftrightarrow Q \preceq \|Q\|_2 I \end{aligned}$$

Finally,  $\sqrt{r} = \sqrt{\sum_{i=1}^r Q_{ii}^2} \leq \|Q\|_2 = \sqrt{\sum_{i=1}^r \sum_{j=1}^r Q_{ij}^2} \leq \sqrt{r^2} = r$  since  $-1 \leq Q_{ij} = \cos(H_i, H_j) \leq 1$ . Therefore, we have:

**Remark 2:**  $I \preceq \nabla^2 f = Q \preceq \|Q\|_2 I$ ; where  $\sqrt{r} \leq \|Q\|_2 \leq r$ .

From Remark 2 setting  $m = 1$  and  $M = \|Q\|_2 \leq r$ , and Remark 1, we have Theorem 4.2.

■

Actually, the exact line search step, from Line 11 to Line 14 in Algorithm 3, guarantees linear convergence of  $\mathcal{O}(1 - \frac{1}{\|Q\|_2})^k$  in the sub-space of passive variables. However, the set



of passive variables changes through iterations. Hence, we employ the fast block coordinate descent steps, from Line 16 to Line 20 in Algorithm 3, that rapidly restrict the domain of solution to converge to the final optimal sub-space of passive variables of the solution. Therefore, the proposed algorithm linearly converges and requires very few iterations.

## 4.2 Complexity

In this section, we analyze the complexity of Algorithm 2 using Algorithm 3 to solve NQP problems. If we assume that the complexity for each iteration contains  $\mathcal{O}(nr^2)$  in computing  $Q = G^T G$ ,  $\mathcal{O}(mnr)$  in computing  $Y = VF^T$ ,  $\mathcal{O}(mr^2)$  in computing  $H = FF^T$ ,  $\mathcal{O}(\bar{k}mr^2)$  in computing  $F$  and  $\mathcal{O}(\bar{k}nr^2)$  in computing  $G$ , where  $\bar{k}$  is the number of iterations, then we have the following Lemma 3:

**Theorem 4.2.** The complexity of each iteration in Algorithm 2 using Algorithm 3 to solve NQP problems is  $\mathcal{O}((m+n)r^2 + mnr + \bar{k}(m+n)r^2)$ . In addition, it is  $\mathcal{O}((m+n)r^2 + rS(mn) + \bar{k}(m+n)r^2)$  for sparse data, where  $S(mn)$  is the number of non-zero elements in data matrix  $V$ .

Theorem 4.2 is significant for big data, because the data is usually big and sparse. In other words,  $mn$  is actually large, but  $S(mn)$  is small; so  $mn \gg (m+n)r^2 + rS(mn) + \bar{k}(m+n)r^2$ . Hence, in experimental evaluation Section 5, we prove that our algorithm can run on large high-dimension sparse datasets such as Nytimes for an acceptable time. In that dataset,  $mnr \gg rS(mn) \gg (m+n)r^2$ , so the running time  $T(m, n, r) \approx rS(mn)$  since  $m, n \gg r$ .

Moreover, Table 2 shows a comparison of the complexity in an iteration of our proposed algorithms (Alo) with other state-of-the-art algorithms' in the literature: Multiplicative Update Rule (MUR) [LS01], Projected Nonnegative Least Squares (PrN) [BBL<sup>+</sup>07], Projected Gradient (PrG) [Lin07b], Projected Quasi-Newton (PQN) [ZC06], Active Set (AcS) [KP08a], Block Principal Pivoting (BLP) [KP08b], Accelerated Hierarchical Alternating Least Squares (AcH), Fast Coordinate Descent Methods with Variable Selection (FCD) [HD11], and Nesterov's Optimal Gradient Method (Nev) [GTY12]. It can be seen that the complexity of our proposed algorithm is highly comparable with that of other algorithms, and the speed of algorithms depend on the number of iterations. In the experimental evaluation, we will show that the iteration number of our algorithm is highly competitive with other algorithms'. Remarkably, moreover, our proposed algorithm has the following properties that other algorithms has yet considered:

- Exploit the sparseness of datasets,
- Runnable for big datasets in limited internal memory systems,
- Convenient to implement in fully paralleled and distributed systems.

**Table 2:** Complexity of an iteration in NMF solvers

Solver	Complexity ( $\mathcal{O}$ )
MUR [LS01]	$mnr + (m + n)r^2$
PrN [BBL <sup>+</sup> 07]	$mnr + (m + n)r^2 + r^3$
PrG [Lin07b]	$(m + n)r^2 + rmn + \bar{k}\bar{t}(m + n)r^2$
PQN [ZC06]	$\bar{k}(mnr + m^3r^3 + n^3r^3)$
BLP [KP08b]	$(m + n)r^2 + mnr + \bar{k}(m + n)r^2$
AcS [KP08a]	$(m + n)r^2 + rmn + \bar{k}(m + n)r^2$
FCD [HD11]	$(m + n)r^2 + rS(mn) + \bar{k}(m + n)r^2$
AcH [GG12]	$(m + n)r^2 + rS(mn) + \bar{k}(m + n)r^2$
Nev [GTLY12]	$(m + n)r^2 + mnr + \bar{k}(m + n)r^2$
Alo	$(m + n)r^2 + rS(mn) + \bar{k}(m + n)r^2$

where  $m, n$  is the matrix size,  $r$  is the number of latent components,  $\bar{k}$  is the average number of iterations,  $\bar{t}$  is the average number of internal iterations, and  $S(mn)$  is the number of non-zero elements of data matrix  $V$ . To easily compare among the algorithms, we consider  $r$  update times for Algorithm FCD as one iteration because the complexity of one update is  $\mathcal{O}(r)$ , while the complexity of one iteration in other accelerated algorithms is  $\mathcal{O}(r^2)$ .

## 5 Experimental evaluation

In this section, we investigate the effectiveness of the proposed algorithm **Alo** by comparing it to 7 carefully selected state-of-the-art NMF solvers belongs to different approaches:

- **MUR**: Multiplicative Update Rule [LS<sup>+</sup>99],
- **PrG**: Projected Gradient Methods [Lin07b],
- **BLP**: Block Principal Pivoting method [KP08b],
- **AcS**: Fast Active-set-like method [KP08a],
- **FCD**: Fast Coordinate Descent methods with variable selection [HD11],
- **AcH**: Accelerated Hierarchical Alternating Least Squares [GG12],
- **Nev**: Nesterov’s optimal gradient method [GTLY12].

**Test cases:** In this experiment, we design two tests using four datasets shown in Table 3. In the first test, 3 typical datasets with different sizes are used: Faces<sup>1</sup>, Digits<sup>2</sup> and Tiny

<sup>1</sup><http://cbcl.mit.edu/cbcl/software-datasets/FaceData.html>

<sup>2</sup><http://yann.lecun.com/exdb/mnist/>

**Table 3:** Dataset Information

Data-sets	$m$	$n$	$r$	MaxIter
Faces	6977	361	60	300
Digits	$6.10^4$	784	80	300
Tiny Images	$5.10^4$	3,072	100	300
Nytimes	$3.10^5$	102,660	100,...,200	300

Images<sup>3</sup>. For these tests, the algorithms are compared in terms of convergence, optimality, and average of the iteration number to investigate their performance and effectiveness. Additionally, average of the the iteration number  $\bar{k}$  for approximate solutions of the sub-problems as NNLS or NQP is to compare the complexity of algorithms. In the second test, a large dataset containing tf-idf values computed from the text dataset Nytimes<sup>4</sup> is used to verify the performance and the feasibility of our parallel algorithms on sparse large datasets.

**Environment settings:** To be fair in comparison, for the first test, the programs of compared algorithms are written in the same language Matlab 2013b, run by the same computer Mac Pro 8-Core Intel Xeon E5 3 GHz RAM 32 GB, and initialized by the same factor matrices  $G_0$  and  $F_0$ . The maximum number of threads is set to 10 while keeping 2 threads for other tasks in the operation system. For the second test, the proposed algorithm is written in Java programming language to utilize the data sparseness.

**Source code:** The source codes of **MUR**, **PrG**, **BIP**, **AcS**, **FCD**, **AcH**, and **Nev** are downloaded from<sup>5</sup>,<sup>6</sup>,<sup>7</sup>,<sup>8</sup>, and<sup>9</sup>. For convenient comparison in the future, we publish all the source codes and datasets in<sup>10</sup>.

## 5.1 Convergence

In this experiment, we investigate the convergence of algorithms by information loss  $\frac{1}{2}\|X - GF\|_2^2$  in terms of time and the iteration number. In terms of time, see Fig. 4, the proposed algorithm Alo is remarkably faster than the other algorithms for the three different-size datasets: Faces, Digits and Tiny Images. Especially, for the largest dataset Tiny Images, the distinction between the proposed algorithm and the runner-up algorithm AcH is easily recognized. Furthermore, in terms of the iteration number, see Fig. 5, the proposed algorithm converges to the stationary point of solutions faster than the others. This observation is clear for large datasets as Digits and Tiny Images. The results are significant in learning NMF models for big data because the proposed algorithm not only converges faster but also uses

<sup>3</sup><http://horatio.cs.nyu.edu/mit/tiny/data/index.html>

<sup>4</sup><https://archive.ics.uci.edu/ml/machine-learning-databases/bag-of-words/>

<sup>5</sup><http://www.cs.toronto.edu/~dross/code/nmf.m>

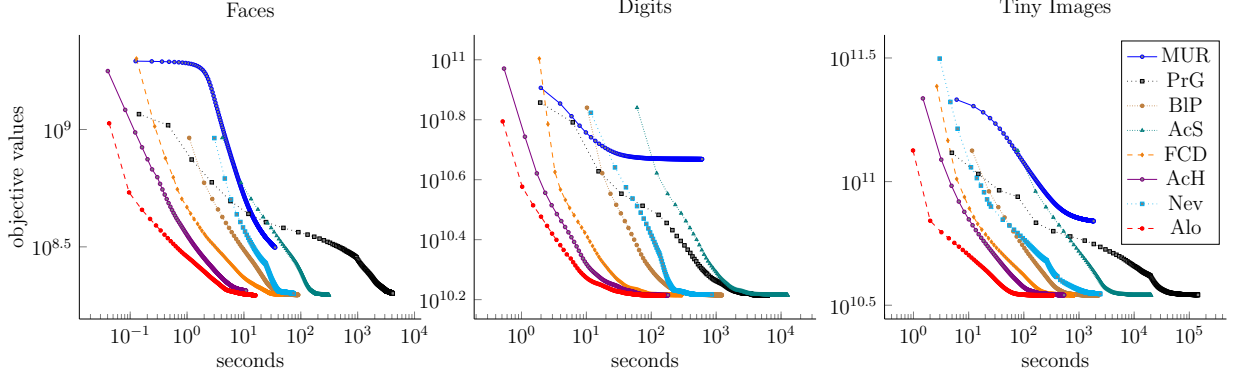
<sup>6</sup><https://github.com/kimjingu/nonnegfac-matlab>

<sup>7</sup><http://www.csie.ntu.edu.tw/~cjlin/nmf/>

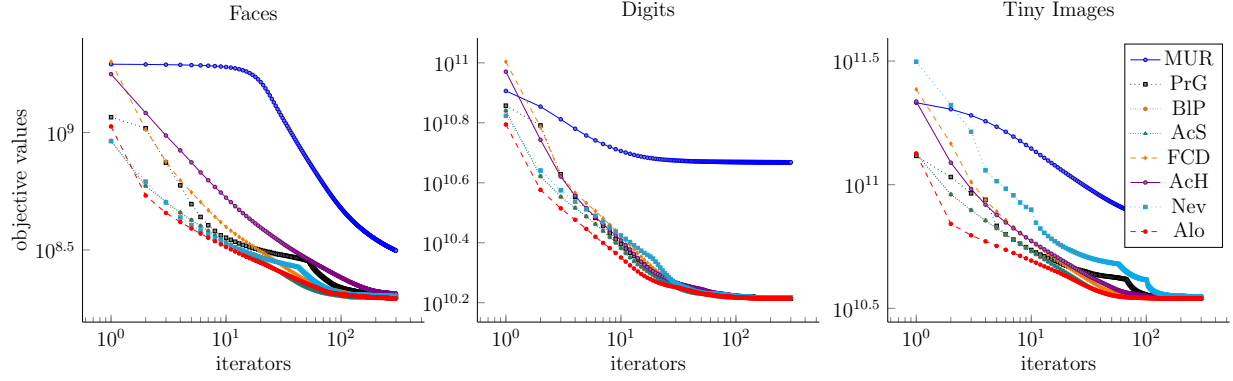
<sup>8</sup>[http://dl.dropboxusercontent.com/u/1609292/Acc\\_MU\\_HALS\\_PG.zip](http://dl.dropboxusercontent.com/u/1609292/Acc_MU_HALS_PG.zip)

<sup>9</sup><https://sites.google.com/site/nmfsolvers/>

<sup>10</sup>[https://bitbucket.org/\[aaa-zzz\]/alnmf](https://bitbucket.org/[aaa-zzz]/alnmf)



**Figure 4:** Objective function values  $\|V - GF\|_2^2/2$  versus CPU seconds for datasets: Faces, Digits, and Tiny Images



**Figure 5:** Objective function values  $\|V - GF\|_2^2/2$  in terms of the iteration number for datasets: Faces, Digits, and Tiny Images

a less number of iterations, and the time of reading and optimization through a big dataset is actually considerable.

## 5.2 Optimality

After more a decade of rapid development, numerous algorithms have been proposed for solving NMF as a fundamental problem in dimension reduction and learning representation. Currently, the difference of the final loss information  $\|V - WH\|_2^2$  among the state-of-the-art methods is inconsiderable in comparison to the square of information  $\|V\|_2^2$ . However, the small difference represents the effectiveness of the optimization methods because NMF algorithms often slowly converge when the approximate solution is close to the optimal local solution. Hence, in Table 4, the final values of the objective function  $\frac{1}{2}\|V - WH\|_2^2$  investigate the optimality and the effectiveness of the optimization methods. Noticeably, Algorithm AcH

**Table 4:** Optimal Values of NMF solvers

Dataset	MUR	PrG	BIP	AcS	FCD	AcH	Nev	Alo
Faces ( $\times 10^8$ )	3.142	2.003	1.975	1.975	1.983	2.058	2.003	<b>1.966</b>
Digits ( $\times 10^{10}$ )	4.659	1.639	1.641	1.641	1.644	1.640	1.646	<b>1.638</b>
Tiny Images ( $\times 10^{10}$ )	6.925	3.483	3.472	3.472	3.474	3.484	3.513	<b>3.468</b>

**Table 5:** Average of Iteration Number  $\bar{k}$ 

Dataset	MUR	PrG	BIP	AcS	FCD	AcH	Nev	Alo
Faces	1.00	321.12	1116.96	102.09	1.54	<b>1.11</b>	29.21	1.29
Digits	1.00	36.70	12503.75	305.94	<b>1.00</b>	1.05	23.36	<b>1.00</b>
Tiny Images	1.00	767.45	12869.12	1086.51	1.38	2.52	29.32	<b>1.27</b>

fast converges over time and has a low average of the iteration number, but it has the optimal values much higher than the proposed algorithm because it uses a time-break technique to interrupt the optimization algorithm. In addition, the proposed algorithm achieves the best optimality for all three datasets. This result additionally represents the robustness of the proposed method, which is highly competitive with the state-of-the-art methods.

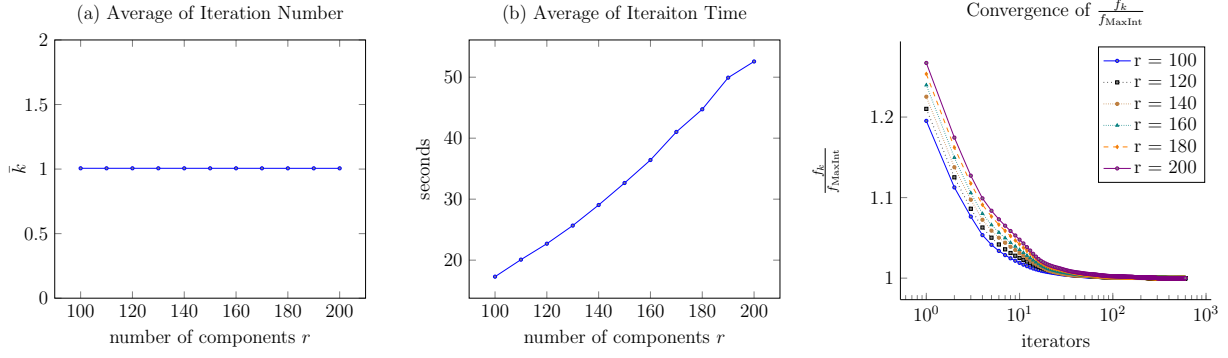
### 5.3 Average of iteration number

In this section, we investigate the complexity of the NMF solvers by average of the iteration number  $\bar{k} = \frac{\text{number of internal iteration}}{\text{MaxIter}(m+n)}$  for approximate solutions of sub-problems as NNLS or NQP because the complexity of algorithms mainly depends on this number, see Table 2. Except for the original algorithm MUR with one update having the worst result, the proposed algorithm Alo employs at least average of the iteration number, see Table 5, especially for large datasets. In addition, the proposed algorithm does not employ any tricks to timely interrupt before one of the stopping conditions is satisfied, while the highly competitive algorithm AcH uses. Therefore, this result clearly represents the fast convergence of Algorithm 3 as it is verified by a large number of NQP problems.

### 5.4 Running on large datasets

In this section, we verify the feasibility of the proposed algorithm in learning NMF model for large datasets. Particularly, the proposed algorithm is implemented by Java programming language to exploit the data sparseness. Additionally, it runs on the large sparse text dataset Nytimes with different numbers of latent components, see Table 3. Interestingly, the proposed algorithm can run with hundreds of latent components by a single computer in an acceptable time.

Fig. 6 shows the performance of our algorithm running on the large sparse dataset Nytimes. Remarkably, the proposed algorithm only uses about 1 iteration on average to sat-



**Figure 6:** average of the iteration number  $\bar{k}$ , average of iteration time, and convergence of  $\frac{f_k}{f_{\text{MaxInt}}}$  in learning NMF model for the dataset Nytimes within the different numbers of latent components

isfy the accelerated condition of approximate solutions. Furthermore, the average of iteration time in learning NMF model linearly increases through the different numbers of latent components. This result totally fits the complexity analysis when  $rn m \gg rS(mn) \gg (m+n)r^2 + \bar{k}(m+n)r^2$ , so the complexity  $T(m, n, r) \approx rS(mn)$  since  $m, n \gg r$ . Additionally, the objective function converges to the stationary point at about the 100<sup>th</sup> iteration within the different numbers of latent components  $r$ , which is the same with the previous datasets.

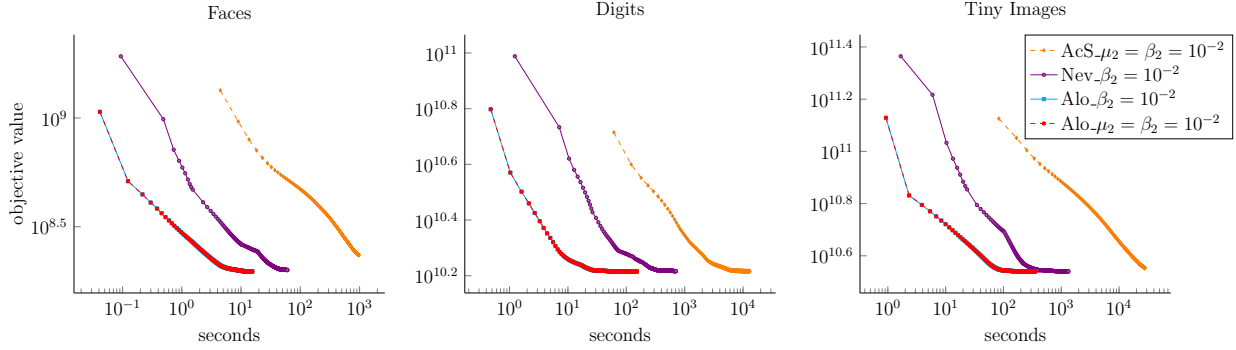
## 5.5 Regularized NMF extensions

In this section, we investigate the convergence of algorithms for regularized NMF extensions on three datasets: Faces, Digits, and Tiny Images. Due to the lack of available codes and the  $L_1$   $L_2$  generalization of the other algorithms, only three algorithms AcS, Nev and Alo are compared within two regularized cases:  $\mu_2 = 10^{-2}$  and  $\mu_2 = \beta_2 = 10^{-2}$ , see Fig. 7. In comparison with other algorithms for regularized NMF extensions, the proposed algorithm Alo converges much faster than algorithms AcS and Nev.

## 6 Conclusion

In summary, our work has two major contributions:

Regarding nonnegative matrix factorization, we propose a flexible algorithm in an unified framework for NMF and its  $L_1$   $L_2$  regularized variants based on full decomposition and a fast combinational algorithm of the anti-lopsided algorithm [NH15] and the greedy coordinate block descent algorithm [HD11]. The proposed algorithm has *linear* convergence rate of  $\mathcal{O}(1 - \frac{1}{r})^k$  in optimizing each matrix factor in the sub-space of passive variables when fixing the other matrix, where  $r$  is the number of latent components. The proposed algorithm is an advanced version of fast block coordinate descent methods and accelerated methods. In theory and practice, the proposed algorithm resolve some current major issues of NMF: fast



**Figure 7:** Convergence of regularized NMF Extensions for algorithms AcS, Nev and Alo within two regularized cases:  $\mu_2 = 10^{-2}$  and  $\mu_2 = \beta_2 = 10^{-2}$

learning algorithm, data sparseness exploit-ability, and parallel distributed feasibility using limited internal memory. Furthermore, the proposed algorithm flexibly adapts with all the variants of  $L_1$   $L_2$  NMF regularizations.

In experimental comparative evaluation, our algorithm overcomes 7 of the most art-the-state algorithms in large datasets about three significant aspects of convergence, average of the iteration number and optimality. In addition, it can fully be parallelized and distributed because the computation using limited internal memory is decomposed into basic computation units as NQP problems. Concerning the feasibility in real applications, the proposed algorithm exploits the data sparseness to learn the huge sparse dataset Nytimes in an acceptable time by a single machine. Finally, the convergence of the proposed algorithm for  $L_1 L_2$  regularized NMF variants is much faster than that of the existing algorithms.

Concerning the optimization techniques for alternating least squares methods, we propose a fast algorithm, Algorithm 3 for NQP problems, which not only has a linear convergence in theory but also is verified in practice about the three significant aspects by a large number of NQP problems conducted inside the NMF framework. Hence, we strongly believe that the algorithm can be effectively employed for alternating least square methods as the key problem in factorization methods.

## Acknowledgement

This work was supported by Asian Office of Aerospace R&D under agreement number FA2386-13-1-4046; and 911 Scholarship from Vietnam Ministry of Education and Training.

## References

- [BBL<sup>+</sup>07] Michael W Berry, Murray Browne, Amy N Langville, V Paul Pauca, and Robert J Plemmons. Algorithms and applications for approximate nonnegative



- matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, 2007.
- [BDJ97] R.Rasmus Bro and S.Sijmen De Jong. A fast non-negativity-constrained least squares algorithm. *Journal of chemometrics*, 11(5):393–401, 1997.
- [Bon11] Silvia Bonettini. Inexact block coordinate descent methods with application to non-negative matrix factorization. *IMA journal of numerical analysis*, 31(4):1431–1452, 2011.
- [Cho08] Seungjin Choi. Algorithms for orthogonal nonnegative matrix factorization. In *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 1828–1832. IEEE, 2008.
- [CP09] Donghui Chen and Robert J Plemmons. Nonnegativity constraints in numerical analysis. In *Symposium on the Birth of Numerical Analysis*, pages 109–140, 2009.
- [CZA07] Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. Hierarchical als algorithms for nonnegative matrix and 3d tensor factorization. In *Independent Component Analysis and Signal Separation*, pages 169–176. Springer, 2007.
- [DS04] D.L. Donoho and V.C. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in neural information processing systems 16: proceedings of the 2003 conference*. MIT Press, 2004.
- [FHN05] Vojtěch Franc, Václav Hlaváč, and Mirko Navara. Sequential coordinate-wise algorithm for the non-negative least squares problem. In *Computer Analysis of Images and Patterns*, pages 407–414. Springer, 2005.
- [GG12] Nicolas Gillis and François Glineur. Accelerated multiplicative updates and hierarchical als algorithms for nonnegative matrix factorization. *Neural computation*, 24(4):1085–1105, 2012.
- [Gil14] N. Gillis. The why and how of nonnegative matrix factorization. *ArXiv e-prints*, January 2014.
- [GNHS11] Rainer Gemulla, Erik Nijkamp, Peter J Haas, and Yann Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 69–77. ACM, 2011.
- [GTLY12] Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan. Nnmf: an optimal gradient method for nonnegative matrix factorization. *Signal Processing, IEEE Transactions on*, 60(6):2882–2898, 2012.

- [GWLT13] Naiyang Guan, Lei Wei, Zhigang Luo, and Dacheng Tao. Limited-memory fast gradient descent method for graph regularized nonnegative matrix factorization. *PloS one*, 8(10):e77162, 2013.
- [HD11] Cho-Jui Hsieh and Inderjit S Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1064–1072. ACM, 2011.
- [Hoy02] Patrik O Hoyer. Non-negative sparse coding. In *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, pages 557–565. IEEE, 2002.
- [Hoy04] Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, 5:1457–1469, December 2004.
- [HV05] Marko Helén and Tuomas Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proc. EUSIPCO*, volume 2005, 2005.
- [KDD13] Sra Suvrit Kim Dongmin and Inderjit S. Dhillon. A non-monotonic method for large-scale non-negative least squares. *Optimization Methods and Software*, 28(5):1012–1039, 2013.
- [KHP14] Jingu Kim, Yunlong He, and Haesun Park. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58(2):285–319, 2014.
- [KP08a] Hyunsoo Kim and Haesun Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30(2):713–730, 2008.
- [KP08b] Jingu Kim and Haesun Park. Toward faster nonnegative matrix factorization: A new algorithm and comparisons. In *Data Mining, 2008. ICDM’08. Eighth IEEE International Conference on*, pages 353–362. IEEE, 2008.
- [KSD06] Dongmin Kim, Suvrit Sra, and Inderjit S Dhillon. *A new projected quasi-newton approach for the nonnegative least squares problem*. Computer Science Department, University of Texas at Austin, 2006.
- [KSD07] Dongmin Kim, Suvrit Sra, and Inderjit S Dhillon. Fast newton-type methods for the least squares nonnegative matrix approximation problem. In *SDM*, pages 343–354. SIAM, 2007.
- [LAW<sup>+</sup>07] Hualiang Li, Tülay Adal, Wei Wang, Darren Emge, and Andrzej Cichocki. Non-negative matrix factorization with orthogonality constraints and its application

to raman spectroscopy. *The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, 48(1-2):83–97, 2007.

- [LH74] Charles L Lawson and Richard J Hanson. *Solving least squares problems*, volume 161. SIAM, 1974.
- [Lin07a] Chih Jen Lin. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *Neural Networks, IEEE Transactions on*, 18(6):1589–1596, 2007.
- [Lin07b] Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.
- [LS<sup>+</sup>99] D.D. Lee, H.S. Seung, et al. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
- [LS01] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.
- [LYF<sup>+</sup>10] Chao Liu, Hung-chih Yang, Jinliang Fan, Li-Wei He, and Yi-Min Wang. Distributed nonnegative matrix factorization for web-scale dyadic data analysis on mapreduce. In *Proceedings of the 19th international conference on World wide web*, pages 681–690. ACM, 2010.
- [Nes83] Yurii Nesterov. A method of solving a convex programming problem with convergence rate  $o(1/k^2)$ . In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.
- [NH15] Duy Khuong Nguyen and Tu Bao Ho. Anti-lopsided algorithm for large-scale nonnegative least square problems. *arXiv preprint arXiv:1502.01645*, 2015.
- [PMCK<sup>+</sup>06] Alberto Pascual-Montano, Jose Maria Carazo, Kieko Kochi, Dietrich Lehmann, and Roberto D Pascual-Marqui. Nonsmooth nonnegative matrix factorization (nsnmf). *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(3):403–415, 2006.
- [PPP06] V Paul Pauca, Jon Piper, and Robert J Plemmons. Nonnegative matrix factorization for spectral data analysis. *Linear algebra and its applications*, 416(1):29–47, 2006.
- [PT94] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.
- [SL01] D Seung and L Lee. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13:556–562, 2001.

- [SLR10] Zhengguo Sun, Tao Li, and Naphtali Rishe. Large-scale matrix factorization using mapreduce. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 1242–1248. IEEE, 2010.
- [TKWB11] C. Thureau, K. Kersting, M. Wahabzada, and C. Bauckhage. Convex non-negative matrix factorization for massive datasets. *Knowledge and information systems*, 29(2):457–478, 2011.
- [WZ13] Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *Knowledge and Data Engineering, IEEE Transactions on*, 25(6):1336–1353, 2013.
- [ZC06] Rafal Zdunek and Andrzej Cichocki. Non-negative matrix factorization with quasi-newton optimization. In *Artificial Intelligence and Soft Computing–ICAISC 2006*, pages 870–879. Springer, 2006.
- [Zha11a] Zhong-Yuan Zhang. Divergence functions of non negative matrix factorization: A comparison study. *Communications in Statistics-Simulation and Computation*, 40(10):1594–1612, 2011.
- [Zha11b] Zhong-Yuan Zhang. Nonnegative matrix factorization: Models, algorithms and applications. 2, 2011.